http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 4 (3): 246-250, 2005
ISSN 1812-5638

Constructing the Virtual Library Data Warehouse from a Blueprint

[1]N. Girija and [2]S.K. Srivatsa
[1]ICFAI Business School, No. 12, Thiruveedian Street, Off. Cathedral Road, Chennai 600086, India
[2]Department of Electronics, Anna University, MIT Campus, Chrompet, Chennai 600044, India

Abstract: Information technology growth has tremendously changed the magnitude of the simple library science to broad information science. Most recent evaluation techniques used in the information science focus mainly on the key areas like user frequencies, requirements of different user groups, predictions regarding the future user requirements and accomplishment of the services. For attaining these, the internet technology helps the technologically based intellectual storage to drive the information science to establish the digital library. The digital library is not just one entity but it is integrated with multiple sources. As a result, building the data warehouse architecture in digital library provides centrally integrated source enabling retrieval of the vast information available at low cost irrespective of the type of format required.

Key words: Digital library, meta data, data warehouse manager, web log, ETL, operational ODS, data mart user information application, bibilomining, member relationship management, service intelligence, acquisition management

INTRODUCTION

Libraries are always places for enhancing the knowledge. According to Dr. S. Ranaganath's narrative principle a library is a growing organization and libraries are moves from the analog world of paper, micro film and vinyl phonograph records to the digital world of computer archives and screen displays. The digital libraries are an exciting change and development of a whole range of underlying theories and technologies construe a paradigm shift. Since, virtual collections from several different sources can be integrated and accessed from anywhere without the user even knowing where the actual sources exist. The digital library is also called the electronic library or the virtual library or the hybrid library. Building a very large database for digital library will always mean management of a terabyte of data in a proven architecture. This study highlights the need for building a suitable data warehousing architecture for the emerging digital library structure.

DATA WAREHOUSE

In 1990, Inmon[1] coined the word data warehouse. The corporate's gave foremost importance only to application-oriented database system, including spatial, temporal, multimedia, active and scientific database, knowledge base and office information base. Heterogeneous database systems and internet based global information system play a vital role and make a huge number of databases and information repositories available for transaction management, information retrieval and data analysis. This technology lead the corporate to think about identifying, storing, managing and retrieving terabyte of information need a narrative model called data warehousing. Data warehousing is a centralized and integrated database for using huge database repository effectively. As a result, applying statistical technique to data warehousing provides the multi dimensional view for analyzing corporate decision making.

A major question here is, how this data warehousing architecture could be made suitable for the emerging trend towards digital library. Digital libraries virtually collect information from different sources integrates them and give access to knowledge without any geographical boundary. There could be the use of computerized indices to online-journal which are accessible from centralized services anywhere, without the user even knowing where the actual source resides. This kind of centralized storage service needs library data warehousing model (Fig. 1).

Major components of the architectures are:

- Staging area
- ETL

Corresponding Author: N. Girija, ICFAI Business School, No. 12, Thiruveedian Street, Off. Cathedral Road, Chennai 600086, India
Tel: 91 9840755112 Fax: 91 044 2811 7674

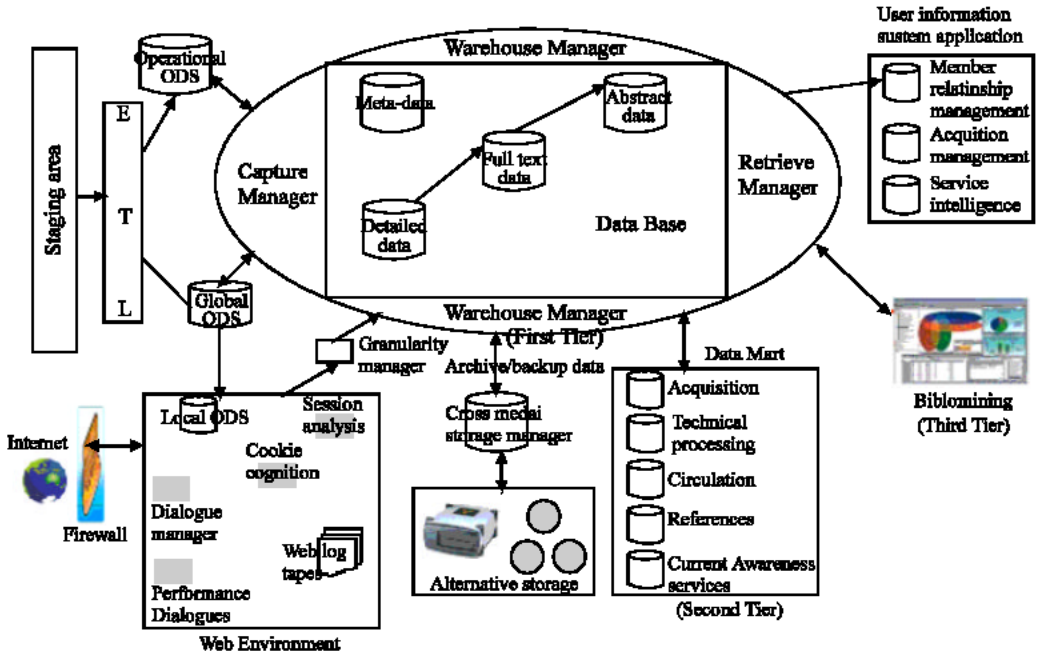

Fig. 1: Virtual library data warehouse architecture

- Operational ODS
- Web environment
- Archived/back up data
- Cross media storage manager
- Alternative storage
- Data mart
- User information application
- Bibilomining

CAPTURE MANAGER

The capture manager is the system component that performs all the operations necessary to support the extracting, cleansing and loading process. Staging areas are needed only where there is a large amount of data to be processed. It is mainly preparing for entering into the ETL setting i.e. ETL is Extraction, Transformation and Loading. The extraction and cleaning process is the key for capture manager for protecting patron privacy during data warehousing. Digitization is the process of conversion of any physical or analogue item into a digital representation or facsimile. Different formats of materials are digitized like:

- Bound volumes, both print and manuscript
- Individual documents
- Photographs, both prints and transparencies
- Microfilms
- Video and audio
- Maps, drawings and other large-format paper items
- Art works
- Textiles
- Physical three-dimensional (3-D) objects

And each format will almost definitely need to be digitized using different methods. The digital capture is only one of the many processes involved in the highly complex sequence of activities that are supported upon the creation, management, use and preservation of digital objects for the long term. The purpose of ETL is to integrate and cleanse data in preparation for entry into the Operational Data Store (ODS). The ODS is fed by the ETL, ODS is a hybrid structure where operational aspects are found mainly for decision support and direct update processing, sometimes not supported in the data warehouse.

The web environment has several components:

- HTML page manager; it interfaces directly with the users on the internet.
- The local ODS; it contains web based data needed for the immediate operation of the web environment.
- Web logs; it is useful for identifying user's interest and particular topic user is frequently visiting and his/her reading habit.
- A web manager; it coordinates different activities that occur inside the web.
- A cookie cognition manager; it checks whether a person has previously been exposed to the web environment.

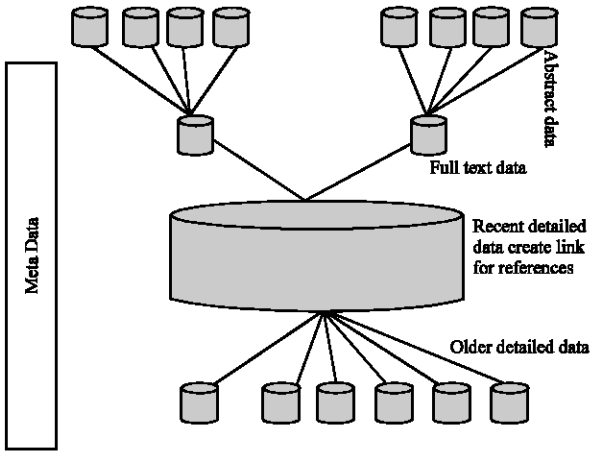

Fig. 2: The structure of data inside the library data warehouse

- A personalization facility; It is the place where different dialogues are personalized for a user who has previously been to the web site.

WAREHOUSE MANAGER

The warehouse manager is the system component that performs all the operations necessary to support the warehouse management[2]. The data found in the data warehouse can be restructured in various ways. Different people need to view the granular data found in the data warehouse in different ways. Designed properly the data warehouse can accommodate all the ways through which data can be accessed.

Data mart is the place where different department place their own data for mainly decision support process. The data mart is unique to a department like (Fig. 2):

- Acquisition; billing, ordering,
- Technical processing; cataloguing , classification
- Circulation; remote login, status check, OPAC access, user requests,
- Reference; e-mail reference, real-time reference, commercial web-based reference
- Current awareness service; recent addition.

The granular data found in the data warehouse are collected and redesigned into an optimal form for the department. Data marts are periodically refreshed from the data warehouse in order to keep their data up to date. When too many data get poured into the data warehouse environment, the volume of dormant data greatly increases. It does not make sense to place all data on high performance disk storage and cannot possibly-economically and technologically - hold all of the data in a data warehouse. Furthermore, with data warehousing there is no need to hold all data in an online mode. Emerging alternative Storage and archival storage is storage that is different from classical disk storage. It lowers the cost, enhances the performance and allows data to reside on a hierarchy of storage.

A granularity manager exists when there is an interface between the data warehouse and the web environment. The granularity manager takes web data-usually click stream data - and condenses the data down to a usable size. The granularity manager edits, deletes, summarizes and otherwise prepares the web based data for analytical processing.

META DATA

Meta data, data about data, is especially used in the context of data that refer to digital resources available across a network. Meta data is linked directly to the resources and so allows direct access to the resource. In other words, the records may contain detailed access information and network addresses. Internet search engines use meta data in the indexing process that they employ to index internet resources. The meta data covers two main functions[3]:

- To provide a means to discover the existence of data set and show how it might be obtained or accessed.
- To document the content, quality and features of a data set, indicating its fitness for use.

The set of data elements for meta data records of document like objects are:

- Title; name given to the resource
- Author; the person or organization primarily responsible for creating the intellectual content of the resource.
- Subject and keywords; keywords or phrases that describe the subject or content of the resource.
- Description; a textual description of the content of the resource.
- Publisher; the entity responsible for making the resource available in its present form.
- Date; the date the resource was made available in its present form
- Resource; resource type, resource format, resource identifier
- Language; the language of the intellectual content of the resource.

- Relation; the relationship of this resource to other resources.
- Right management; a link to a copyright notice, statement terms of access to the resource.

Meta data provide multiple pathways for finding each item and render each item in the collection uniquely identifiable are very important. There are different types of meta data:

- Descriptive meta data: it refers to the attributes of the object like title, subject, date,keyword etc.
- Structural meta data: it describes the structure and relationship of a set of digital objects.
- Administrative meta data: it describes the capture process which is needed to manage digital object throughout the whole of its lifecycle like initial capture setting, file formats, data of capture and compression.

RETRIEVE MANAGER

The retrieve manager is the system component and the eventual target is to enhance the capture manger and warehouse manager operations from information management to knowledge management. The value of information is not quantifiable and it is based on how influencing it is in attaining knowledge and using that to improve the user service. From the user's perspective, this can empower the user with real actionable knowledge in solving their real information problems. The key functions of retrieve manager system are to establish a fundamental infrastructure for decision-making, establish relationship with user's community, provide services associated with user's information and learning and professional effectiveness via bibliomining.

MEMBER RELATIONSHIP MANAGEMENT (MRM)

Information is a basic resource for all human beings. Save the time of the reader Dr. S. Ranganath's principle replicates the significance of Member Relationship Management (MRM). MRM by definition is about managing the entire relationship between the member and library, not just knowing where they are located and what people work there. Ensuring accurate member information is not an easy task. Members are dynamic, as are their information. In addition to servicing member, member must want to use member relationship management system and then only it adds the contribution value to the information collected about member. In the digital environment, the information profession is the user's hero and probably the only group intent on simplifying access to and retrieval of information form this glut of resources. MRM system designed must;

- Understand ways of categorizing users like novice, expert, occasional, frequent, child, older, adult and user with special needs.
- Be aware of the different types of dialogue styles through which search strategies may be executed
- Be aware that the users have different types of search patterns
- Provide the value of user models and cognitive modeling.

ACQUISITION MANAGEMENT

Acquisition is price verification; bibliographic details and downloading of bibliographic records, billing and ordering. Foremost business of libraries is to collect and preserve the materials published. Purchasing digital content[4] suggests that information resources may be offered to a library in one of three main marketing modes:

- The site license: The site license is offered to the organization as a whole and through the library directly to the division most likely to use.
- The individual subscription: Directly analogous with print subscription but the individual loses out on low counts: they miss the technical and subject guidance of the library plus they have access to the materials but not any ownership in the longer term.
- The pay-per-view option: The individual may access the resource to a certain level at zero fee or low-cost subscription of the purposes of identifying the required resource. Once resource discovery has occurred the user then pays a fee for full use of the item, such as the printing and viewing of the full text of an article and licensed use or printing of a higher-resolution.

Acquisition management also concentrates on full bibliographic information in file structure. Most of the files are linked with digitize images and OCR texts that made up each article. Articles are linked into issues into volumes, journal titles and finally all of these are compiled into a data set.

SERVICE INTELLIGENCE

The data warehouse incorporates data from all sources and supports multiple library functions. It is

accessed by many departments in the library. Service Intelligence (SI) is the gathering and analysis of vast amounts of data in order to gain insights that derive strategic and tactical decisions which, in turn, will improve performance in the library service. Dr. S. Rangathan's five laws, the major focus of the three laws are documents are for use, every reader reads his or her document and every document has its reader. They mainly concentrate on how do utilize each and every document or object (in digital library) to provide excellent service for every member. Service intelligence helps track of what really works and what doesn't. Service intelligence data are collected from internal sources, such as web log data, session analysis, cookie recognition and ODS. Data can be related to all facets of the service, such as user request, user transactions, user enquiry, indexing services, referral service like email, real-time-chat technology and commercial web-based, current awareness service like journal tables of contents, recent addition, conference and Collaborative Digital Reference Service (CDRS)[5].

BIBLIOMINING

In the digital library environment, facing the need to adapt to the new climate of the WWW not only create an information explosion but make unstructured and unrestricted information dissemination and retrieval. In virtual reference desk service, the challenge in new information technologies is to assist in organizing, investigating, retrieving and also predicting future user needs and this is called bibliomining. The term Bibliomining was first coined by Nicholoson[6] in discussing data mining for libraries. Bibliomining is the combination of data mining, bibliometrics, statistics and reporting tools used to track patterns in authorship, citation and extract patterns of behaviour. The biblioming aids decision making for librarians and can better predict the demand for new items in order to determine how many copies of a work are to be ordered. Different bibliomining techniques like logistic regression, non-parametric discriminate analysis, classification trees and neural networks are applied to the data in order to determine the best set of criteria to discriminate between scholarly research and other studies.

CONCLUSIONS AND RECOMMENDATION

The main aim of the research was to suggest a data warehousing model which suits for emerging web environment in digital library. Lot of future research is possible in this model. Mainly various approaches of meta data could be extended for data mart and ODS. Here only a few user application systems like member relationship management, acquisition management and service intelligence are proposed. To extend further more user application systems are to be identified like Annotation Management System, Ontology-based Multilingual System, Knowledge Organization Systems[7] and Digital Asset Management (Dam) system. Wide-area file system is a large repository of digital objects that requires both storage and content management like Andrew File System, Semantic File System (SFS) extent to support more complex link analysis, query processing and retrieval. Data warehousing converges to growing open source software movement called super data warehousing which is to accomplish new levels of performance, scalability and cost effectiveness.

REFERENCES

1. Inmon, W.H., 2001. Building the Corporate Information Factory from a Blueprint. Part III. Billinmon.com
2. Anahory, S. and M. Dennis, 2003. Data Warehousing in the Real World: A Practical Guide for Building Decision Support Systems. New Delhi: Pearson Education (Singapore) Pte. Ltd., pp: 9-10.
3. Dillon, M., 2000. Meta data for web resources: How meta data works on the web. In: Bicentennial Conference on Bibliographic Control for the New Millennium: Confronting the Challenges of Networked Resources and the Web, Library of Congress, Washington, DC, http://lcweb.loc.gov/catdir/bibcontrol/dillon.html
4. Getz, M., 1997. Evaluating digital strategies for storing and retrieving scholarly information. J. Lib. Admn. XXIV: 81-98. Also appears in Sul H. Lee, Ed. Economics of Digital Information: Collection, Storage and Delivery, (New York). Haworth Press, 1997, pp: 81-98.
5. Deegan, M. and T. Simon, 2002. Digital Futures: Strategies for the Information Age. London: Library Association Publishing, pp: 131-134.
6. Nicholson, S., 2003. Bibliomining for automated collection development in a digital library setting: Using data mining to discover web-based scholarly research work. J. Am. Soc. Inform. Sci. Technol., 54: 1081-1090.
7. Rowley, J.E. Jr., 2004. Organizing Knowledge : An Introduction to Managing Access to Information (3rd Edn.), England: Ashgate publishing Limited, pp: 12-13.